Segmentation Parole/Musique pour la transcription automatique

Joseph Razik, Dominique Fohr, Odile Mella, Nathalie Parlangeau-Vallès

Segmentation Parole/Musique pour la transcription automatique

J. Razik,[1]D. Fohr[1], O. Mella[1] et N. Parlangeau-Vallès[2]

(1) LORIA-CNRS-INRIA-UHP
Campus Scientifique BP 239 54506 Vandœuvre-lès-Nancy Cedex, France
Tél.: ++33 (0)3 83 59 30 00 - Fax: ++33 (0)3 83 27 83 19
Mél: {fohr, mella, razik}@loria.fr
(2) UNIVERSITE TOULOUSE I
21 allée de Brienne– 31000 Toulouse, France
Mél: nathalie.valles@univ_tlse1.fr

ABSTRACT

The speech/music segmentation process is very useful as a first step for different tasks like speech recognition or automatic transcription. In this article, we present some studies about the use of MFCC for this speech/music segmentation. We mainly use a real-world broadcast corpus with various backgrounds and superimposed segments (speech with music). We investigate the role of the number of cepstral coefficients, the influence of different kinds of dynamic parameters, and the robustness of some of them when a mismatch between train and test conditions occurs. So we can notice that the standard MFCC coefficients with the first and second derivatives achieve good results. But, better performances were obtained with dynamic parameters and mainly with the variance of the static coefficients computed on a long–term window (1s).

1. INTRODUCTION

La segmentation du signal audio en segments de parole et de musique est une étape indispensable dans le traitement automatique des documents audiovisuels. En effet, dans des applications comme la transcription automatique ou la recherche de mots-clés dans les documents audiovisuels, il faut éviter d'activer un système de reconnaissance « grand vocabulaire » sur une portion de signal correspondant uniquement à de la musique ou à des chansons [1]. De même, dans le cadre de l'indexation d'un document sonore, il est important de repérer l'alternance de segments de parole, de musique ou de jingles car celle-ci correspond à une certaine structure du document (bulletin d'information, publicité,…)[2].

Plusieurs travaux de recherche ont été menés dans le domaine de la segmentation Parole/Musique aussi bien pour tester différents types de classifieurs [3],[4] que pour déterminer les meilleurs paramètres acoustiques permettant de discriminer un segment de musique par rapport à un segment de parole [3],[5]. Nous avons souhaité compléter ces études en nous plaçant dans le domaine de la transcription automatique d'émissions radiophoniques ou télévisuelles, c'est-à-dire :

- d'une part, étudier la pertinence de paramètres directement issus des coefficients MFCC qui ont l'avantage d'être déjà calculés en vue de faire la reconnaissance de parole ;
- d'autre part, découper en segments, de Parole, Musique et Parole/Musique, des corpus de tests réels, radiophoniques de longue durée (plusieurs dizaines de minutes), de contenus très variables avec notamment des fondus-enchaînés entre la parole et la musique instrumentale ou chantée.

Après avoir présenté notre système de segmentation, nous décrirons les spécificités des corpus utilisés puis présenterons les résultats et les conclusions des différentes expérimentations.

2. LE SYSTÈME DE SEGMENTATION

2.1. Paramétrisation

Notre paramétrisation étant fondée sur les coefficients cepstraux, le signal audio est échantillonné à 16 kHz et les coefficients MFCC (Mel-Frequency Cepstral Coefficients) sont calculés à partir d'un banc de 24 filtres Mel appliqué toutes les 10 ms sur des fenêtres de 32ms.

2.2. Modélisation

Notre segmentation Parole/Musique repose sur la mise en compétition de quatre modèles constitués de mélanges de 32 gaussiennes (GMM, Gaussian Mixture Models) :

- Parole (P),
- Musique instrumentale (MI),
- Chansons (MC),
- Parole-Musique (P&M), parole sur fond de musique instrumentale ou de chansons.

Pour des raisons pratiques, ces GMMs ont été codés sous la forme de modèles de Markov cachés à un état dont l'apprentissage a été réalisé à l'aide de la boîte à outils HTK [6].

2.3. Le processus de décision

La mise en compétition directe trame par trame des quatre modèles présentés ci-dessus peut conduire à des segments de parole ou de musique ayant une durée minimale irréaliste de 10 ms. Aussi avons-nous décidé d'imposer une durée minimale d'une seconde pour chaque segment reconnu, directement codée dans le modèle. Comme le montre la Figure 1, chacune des classes de segments recherchée est donc modélisée comme la succession de 100 modèles HMM à un état.

L'alignement sur le signal audio, par l'algorithme de Viterbi, de la meilleure séquence des modèles en compétition fournit la segmentation en segments de Parole,(P), Musique (MI ou MC) ou Parole sur Musique (P&M).

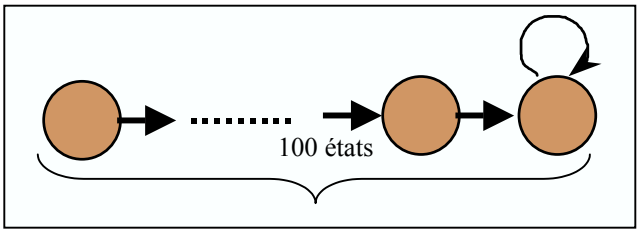

Figure 1: La topologie de chaque modèle.

3. LES CORPUS

3.1. Le corpus radiophonique réel

Ce corpus est constitué de six émissions radiophoniques francophones et hispanophones de 20 à 60 minutes chacune. Les enregistrements ont été ré-échantillonnés à 16 kHz mono et étiquetés manuellement en Parole, Musique Instrumentale, Chansons et Parole-Musique dans le cadre du projet RAIVES [2]. Ces enregistrements comprennent aussi bien des bulletins d'informations que des émissions thématiques avec interviews ou des programmes musicaux.

Ce corpus réel est un corpus difficile. En effet, ces enregistrements contiennent des interventions de locuteurs très différents (hommes, femmes, enfants, de nationalités variées) dans lesquelles la parole large bande enregistrée en studio alterne avec de la parole téléphonique. Certaines interviews ont été réalisées dans un environnement bruité (ambiances de rue et de café, « cocktail party », traduction simultanée). Ces émissions comportent également de nombreux segments où la parole est superposée à de la musique instrumentale ou à des chansons. Cette superposition s'effectue avec une variation continue du rapport de puissance entre la musique et la parole (fondu-enchaîné).

3.2. Le corpus « Scheirer »

Le corpus que nous appelons « Scheirer » est le corpus utilisé par E. Scheirer et M. Slaney dans leur étude [3]. Il est constitué de 80 morceaux de 15 secondes de parole presque exclusivement anglaise et de 100 morceaux de 15 secondes de musique instrumentale ou de chansons extraits de la radio (soit un total de 45 minutes). Ce corpus ne comporte pas de parole sur musique et chacun des morceaux est homogène. Toutefois, après écoute du corpus, nous avons ré-étiqueté 2 fichiers de parole (P) en Parole-Musique (P&M). Ce corpus comporte un éventail plus vaste de styles de musiques et de chansons (jazz, pop, country, …) que notre corpus radiophonique. Il contient également de la parole large bande et téléphonique.

3.3. Les corpus utilisés pour le test

Les tests de segmentation ont été effectués sur :

- le corpus « Scheirer », anglais: 97%, espagnol:3%,
- sur l'émission, appelée « Cigarette » de 24 minutes extraite du corpus réel, entièrement en français,
- sur le corpus « Croisé » constitué des 6 fichiers du corpus réel testés en validation croisée, soit environ 3 heures de données, français : 64%, espagnol : 36%.

La Table **1** fournit les proportions des différents types de segments dans les 3 corpus de test.

Table 1 : Composition en % des trois corpus de test**.**

	Musique	**Parole-Musique**	**Parole**
Cigarette	14	10	76
Croisé	9	15	76
Scheirer	56	0.6	43.4

3.4. Les corpus utilisés pour l'apprentissage

Nous avons appris nos quatre modèles sur deux corpus : le corpus réel radiophonique et des CDs de musique instrumentale et de chansons. La Table **2** donne la taille moyenne des données d'apprentissage pour chacun des modèles.

Table 2 : Corpus d'apprentissage pour chaque modèle.

Modèle	**P**	**MI**	**MC**	**P&M**
Corpus radio réel	136mn	11mn	6mn	26mn
CDs audio	_	30mn	90mn	_

Le contenu du corpus radio réel varie légèrement selon les tests. En effet, pour les tests portant sur « Cigarette » et sur « Scheirer », les modèles ont été appris sur le même corpus d'apprentissage : les 5 autres émissions du corpus radiophonique réel. Alors que pour le test « Croisé », le corpus d'apprentissage comprend 5 fichiers parmi 6.

4. EXPÉRIMENTATIONS

4.1. Méthodologie d'évaluation

Lors de la phase de test, les fichiers de test sont segmentés en segments de Parole (P), Musique (MI ou MC) et Parole-Musique (P&M). L'évaluation est faite trame par trame (10ms). Nous avons choisi des taux d'erreur reflétant au mieux l'objectif d'une telle segmentation, placée en amont d'un système de reconnaissance. Elle doit éliminer les segments de musique, ne pas éliminer les segments contenant de la parole et classer les segments de parole sur musique à part. En effet, on pourra les traiter différemment lors de la phase de reconnaissance de parole : appliquer une méthode de compensation ou bien les reconnaître avec d'autres modèles phonétiques. Nous avons donc calculé les deux taux d'erreur suivant:

- **EG** : le taux d'erreur global, somme des pourcentages des trames de type X non reconnues comme X, X∈{Parole, Musique, Parole-Musique},
- **ET** : le taux d'erreur pour la transcription automatique, qui est aussi le taux d'erreur Parole/Non-parole, c'est la somme des pourcentage des trames :
  - de Musique reconnues comme Parole,
  - de Musique reconnues comme Parole-Musique,
  - de Parole reconnues comme Musique,
  - de Parole-Musique reconnues comme Musique.

Ces taux d'erreur ont été calculés sur les 3 corpus de test avec un intervalle de confiance de ± 0,2%.

4.2. Nombre de paramètres cepstraux

Les premiers tests ont été effectués en combinant les dérivées du premier (Δ) et du deuxième ordre (ΔΔ) aux coefficients statiques C_0 à C_N avec N∈{8,11,15}, ce qui conduit à 27, 36 ou 48 paramètres. La Table 3 montre les taux d'erreur obtenus sur les 3 corpus**.**

Table 3 : Influence du nombre de MFCCs.

	Cigarette		Croisé		Scheirer	
%	**ET**	**EG**	**ET**	**EG**	**ET**	**EG**
27 MFCC	6.2	32.5	4.1	25	1.8	13.7
36 MFCC	5.31	28.4	3.9	20.2	1.8	6.6
48 MFCC	5.7	28.9	4.3	21.9	1.8	9.2

Les résultats sont meilleurs mais moins pertinents sur le corpus « Scheirer », quel que soit le nombre de coefficients. Ceci s'explique par le fait qu'il n'y a pratiquement pas de superposition parole-musique dans ce corpus (0,6%). De plus, nous pouvons observer qu'il est nécessaire de modéliser finement les différences entre parole et musique en utilisant un nombre suffisant de coefficients cepstraux (12). En revanche, trop de coefficients nuit à la segmentation, les coefficients supplémentaires ne semblent pas modéliser d'information utile pour discriminer la parole de la musique. Par ailleurs, la suppression du coefficient statique C0 (35 coefficients) donne des performances légèrement moins bonnes sur « Cigarette » et identiques sur « Scheirer ».

4.3. Pertinence des paramètres dynamiques

Afin d'étudier la pertinence des paramètres dynamiques par rapport aux paramètres statiques nous avons comparé les résultats de cinq paramétrisations :

- MFCC classique : 12 C_i + 12 Δ+ 12ΔΔ,
- uniquement les 12 coefficients statiques,
- uniquement les 12 dérivées du 1er ordre,
- uniquement les 12 dérivées du 2e ordre,
- uniquement les variances des 12 coefficients statiques calculées sur 1s.

Table 4 : Influence des paramètres dynamiques.

	Cigarette		Croisé		Scheirer	
%	**ET**	**EG**	**ET**	**EG**	**ET**	**EG**
12 C_i + 12 Δ+ 12ΔΔ	5.3	28.4	3.9	20.2	1.8	6.6
12 Ci	17.7	32.5	9.6	31.7	3.5	8.8
12 Δ	2.5	26.0	5.1	27.1	2.2	9.3
12 ΔΔ	2.5	24.1	5.2	26.9	2.8	9.8
Variance des C_i	3.5	27.4	3.0	19.8	1.7	3.7

Les taux d'erreur présentés dans la Table 4 sont cohérents pour les corpus « Croisé » et « Scheirer » et montrent l'importance des paramètres dynamiques comme les dérivées du premier ordre. En effet, la parole étant acoustiquement plus variable à moyen terme que la musique (alternance de voyelles et de consonnes), les paramètres dynamiques devraient mieux discriminer ces deux phénomènes.

Par ailleurs, cette discrimination est meilleure lorsque cette dynamique est calculée à long terme : variance sur une seconde. Ces résultats corroborent ceux de E. Scheirer et M. Slaney qui ont trouvé que les variances de leurs paramètres acoustiques étaient plus discriminantes que les paramètres eux-mêmes [3]. Les résultats légèrement différents sur « Cigarette » sont sans doute liés au contenu spécifique de cette émission.

Les bonnes performances obtenues par les paramètres dynamiques sur nos corpus radiophoniques peuvent également s'expliquer par le fait que l'effet du canal de transmission est fortement atténué par ceux-ci. En effet, notre modèle de Parole modélise globalement la parole en studio, la parole téléphonique et la parole bruitée.

Cette conclusion devrait rester valable lorsque il y a une forte différence entre les conditions d'apprentissage et de test. C'est ce que nous avons testé dans le paragraphe suivant.

Influence des paramètres dynamiques avec un corpus d'apprentissage de laboratoire

Pour cette expérimentation, nous avons construit un corpus d'apprentissage de laboratoire. Le modèle de Parole (P) a été appris à partir de phrases lues du journal « Le Monde » (45 minutes) auxquelles a été ajouté le résultat du filtrage dans la bande téléphonique de ces phrases. Les modèles de Musique instrumentale (MI) et de Chansons (MC) ont été appris sur les Cds.

Pour apprendre le modèle Parole-Musique (P&M), nous avons élaboré un corpus de parole comportant de la musique ou des chansons en bruit de fond en mixant les deux corpus précédents en utilisant trois valeurs de rapport de Parole/Musique : 5, 10 et 15 dB.

Aucune donnée radiophonique n'a donc servi à l'apprentissage des quatre modèles. Le test de segmentation a été réalisé sur les corpus « Cigarette », « Scheirer » et sur l'intégralité du corpus radiophonique réel, appelé « Radio » dans la Table 5.

Table 5 : Taux d'erreur avec un corpus d'apprentissage sans données radiophoniques

Erreur **G**lobale **(%)**	**Cigarette**	**Radio**	**Scheirer**
36 MFCC	51.6	34.5	21.9
12 Δ	27.6	24.9	9.2
Variance des Ci sur 1s	27.3	20.7	9.2

Avec les paramètres standard, comme on pouvait s'y attendre, les performances sont bien plus mauvaises qu'avec un corpus d'apprentissage radiophonique. En revanche, les dérivées de premier ordre et surtout la variance des coefficients statiques sont des paramètres robustes au changement de condition. En effet, avec ces coefficients dynamiques on atteint pratiquement les mêmes taux de segmentation que ceux présentés dans la Table 4. Notons que la segmentation réalisée avec la paramétrisation standard ($12\ C_i + 12\Delta + 12\Delta\Delta$) sur un corpus de test similaire au corpus de laboratoire donne un taux EG de 0%, sauf dans le cas où le rapport musique à bruit est de 5dB, pour lequel EG est de 3%.

5. DISCUSSION ET CONCLUSION

Nous avons réalisé un système de segmentation parole/musique segmentant un signal audio en Parole, Musique ou Parole sur fond musical. Avec une paramétrisation standard MFCC ($12\ C_i + 12\Delta + 12\Delta\Delta$), il donne de bons résultats même sur un corpus radiophonique réel et difficile si les conditions entre les corpus d'apprentissage et de tests ne sont pas trop différentes. Seules 4% des trames sont mal classées en vue de leur traitement ultérieur par un système de reconnaissance (trames de parole rejetées ou trames de musique transmises au reconnaisseur). Ces résultats portent sur 3 langues : anglais, français et espagnol et sur une grande variété de musiques. Rappelons que M. Carey et al [4] ont montré la pertinence des coefficients MFCCs mais sur des segments homogènes de 10s de parole multilingue et de musique, issus de bases de données classiques en traitement de la parole.

Les paramètres dynamiques comme les dérivées du premier ordre et surtout la variance sur 1s des 12 C_i obtiennent de très bonnes performances aussi bien sur le corpus radiophonique que sur le corpus « Scheirer » : respectivement 80,2% et 96,3% de trames correctement étiquetées. L'écart entre les deux taux d'erreur ET et EG dans les tableaux 3 et 4 est dû en général pour 25% des cas à la confusion de la parole bruitée en Parole-Musique et pour 75% à la confusion parole sur fond musical en Parole. Ceci est sans doute dû à l'étiquetage manuel ; en effet, même si la musique en arrière-plan est très faible, le segment a été étiqueté comme Parole-Musique.

Les paramètres dynamiques sont également robustes lorsque les corpus de test et d'apprentissage sont très différents. Nous ne pouvons pas comparer nos résultats avec ceux de E. Scheirer et M. Slaney [3] car les corpus d'apprentissage ne sont pas identiques et leur corpus de test ne comporte pas de parole sur fond musical. E. Scheirer et M. Slaney n'ont pas testé la pertinence de la paramétrisation MFCCs mais ils ont mis en évidence d'autres paramètres acoustiques comme la modulation de l'énergie à 4Hz ou la variance de la dérivée de l'amplitude du spectre. Nous nous proposons donc dans une étude future de coupler ces paramètres avec la variance des coefficients statiques Ci et de les valider sur nos corpus réels.

REMERCIEMENTS :

- Ce travail a pu être mené grâce au projet CNRS STIC-SHS RAIVES.
- Nous remercions Messieurs Scheirer et Slaney de nous avoir aimablement fourni leur corpus.

BIBLIOGRAPHIE

[1] J-L. Gauvain, L. Lamel and G. Adda, "Partitioning and transcription of broadcast news data", In *Proceedings of ICSLP*, 1998.

[2] N. Parlangeau-Vallès et al. "Audio Indexing on the Web: A preliminary study of some audio descriptors", In *Proceedings of SCI*, 2003.

[3] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", In *Proceedings of ICASSP*, pages 1331-1334, 1997.

[4] J. Pinquier, C. Sénac and R. André-Obrecht, "Speech and Music classification in audio documents", In *Proceedings of ICASSP*, 2002.

[5] M.J. Carey, "A comparison of features for speech, music discrimination", In *Proceedings of ICASSP*,1999.

[6] SJ. Young and al., "*The HTK Book***"**, Cambridge, England ,Entropic Ltd., 1995.